

A multi-resolution model to capture both global fluctuations of an enzyme and molecular recognition in the ligand-binding site

Aoife C. Fogarty[†], Raffaello Potestio[†], Kurt Kremer^{*†}

[†] *Max Planck Institute for Polymer Research, Ackermannweg 10, 55128 Mainz, Germany*

^{*} *Corresponding author, kremer@mpip-mainz.mpg.de*

This is the pre-peer reviewed version of the following article:

Fogarty, A. C., Potestio, R. and Kremer, K. (2016), A multi-resolution model to capture both global fluctuations of an enzyme and molecular recognition in the ligand-binding site. *Proteins*. doi: 10.1002/prot.25173,

which has been published in final form at <http://onlinelibrary.wiley.com/doi/10.1002/prot.25173/full>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

Abstract

In multi-resolution simulations, different system components are simultaneously modelled at different levels of resolution, these being smoothly coupled together. In the case of enzyme systems, computationally expensive atomistic detail is needed in the active site to capture the chemistry of substrate binding. Global properties of the rest of the protein also play an essential role, determining the structure and fluctuations of the binding site; however, these can be modelled on a coarser level. Similarly, in the most computationally efficient scheme only the solvent hydrating the active site requires atomistic detail. We present a methodology to couple atomistic and coarse-grained protein models, while solvating the atomistic part of the protein in atomistic water. This allows a free choice of which protein and solvent degrees of freedom to include atomistically, without loss of accuracy in the atomistic description. This multi-resolution methodology can successfully model stable ligand binding, and we further confirm its validity via an exploration of system properties relevant to enzymatic function. In addition to a computational speedup, such an approach can allow the identification of the essential degrees of freedom playing a role in a given process, potentially yielding new insights into biomolecular function.

1 Introduction

Biomolecules in solution are complex, heterogeneous systems with length and timescales covering many orders of magnitude.^{1–3} Simulating such systems therefore often leads to requirements that are difficult to reconcile, namely (i) large systems and long simulation times, and (ii) accurate models that contain sufficient physical and chemical detail to describe a given phenomenon, and that are therefore computationally expensive. As such, biomolecular simulations may benefit from a concurrent multi-resolution approach. This involves identifying those parts of the system where the physical and chemical detail plays an essential role in the phenomenon of interest, and describing them using a sufficiently high-resolution model, while using a less detailed, computationally more efficient model for the remainder of the system.^{4,5} Conversely, instead of merely obtaining computational speedup, the goal may in fact be to identify precisely which degrees of freedom play a role in a biomolecular process, via a model in which the degrees of freedom included at a certain resolution can be arbitrarily varied.

In most concurrent multi-resolution simulation approaches, each distinct system component, such as protein, nucleic acid, lipid membrane, or aqueous solvent, is modelled in its entirety using one level of resolution, e.g. an atomistic protein in a coarse-grained solvent or embedded in a coarse-grained membrane,^{6–8} or a coarse-grained protein in a continuum environment.⁹ However, when the goal is to construct a model which includes only the minimum possible number degrees of freedom, one must be able to place boundaries between resolutions at any arbitrary place within the system. Take for example the case of enzymatic function. In its simplest form an enzyme can be seen as being composed of two parts: an active site, at which the substrate binding and catalytic reaction occur, and the remainder of the enzyme, which exists to induce in the active site those properties necessary for its function; these may include its structure, conformational fluctuations, dynamics, or electric field.^{10,11} Furthermore, the aqueous solvent is known to play an essential role in enzymatic function.^{12,13} An accurate model of substrate binding therefore requires at minimum an atomistic level of detail in the description of the substrate, binding site and neighbouring water molecules. However, the global structure and conformational fluctuations of a protein can be captured on a more coarse-grained level,^{14,15} as can the aqueous solvent further away from the binding site.^{16,17}

Here we propose a new approach which allows one to model at an atomistic resolution only the precise subset of degrees of freedom really necessary for the study of a given phenomenon, even when this leads to a boundary between resolutions which falls within a macromolecule, or includes only part of the solvent. For the enzyme system studied here, this means a coarse-grained protein in which is embedded an atomistic binding site, solvated by a small sphere of atomistic water, which freely exchanges with the reservoir of coarse-grained water filling the remainder of the simulation box. In the construction of such a multi-resolution model, two methodological issues arise: (i) the coupling between particles with different resolutions connected by bonds, and (ii) how to describe solvent particles diffusing between regions at different levels of resolution. The former issue has been explored in a small number of polymer^{18,19} and biomolecular^{20–23} studies. In particular, Neri and co-workers developed a model in which an atomistically detailed active site was incorporated into a coarse-grained Gō model.²⁰ Here, we present a method for inserting atomistic residues into a protein whose essential structure and conformational fluctuations are described using an Elastic Network Model (ENM). We note that the approach presented here differs from that of Neri et al.²⁰ both in our treatment of the atomistic-coarse-grained coupling within the protein, and in our inclusion of explicit solvent. The second methodological issue, that of allowing solvent particles to change their resolution on the fly, can be tackled using the Adaptive Resolution Scheme (AdResS),²⁴ in which interpolation of atomistic and coarse-grained forces across a transition region allows for a smooth coupling between different resolution levels, and free diffusion of solvent particles throughout the simulation box. Only the small subregion of water solvating the atomistic protein active site is then described at an atomistic level, but behaves as though it were in a fully atomistic system.^{25–28}

This multi-resolution approach not only leads to greater computational efficiency via both a reduction in the number of degrees of freedom simulated, and via the improved efficiency in the sampling of slow, large-scale protein fluctuations which is inherent in the use of a coarse-grained protein model. It can also allow the simulation of large biomolecular systems where atomistic structure is not known everywhere and where atomistic simulations are not even possible, but where low-resolution experimental data can still be exploited to parametrise coarse-grained models

for some parts of the system.

2 Methods

We demonstrate our methodology on an aqueous solution of hen egg-white lysozyme (HEWL), a widely studied 14 kD enzyme that hydrolyses glycosidic bonds in polysaccharides. In our model, the ligand and binding site are represented in atomistic detail. The precise set of protein residues modelled atomistically is determined for a given ligand by the specific H-bonding and hydrophobic contacts between that ligand and the binding site. The protein model is not adaptive, i.e. the resolution of a given residue is *either atomistic or coarse-grained* and this does not change during the simulation. Solvent molecules, in contrast, may diffuse towards or away from the binding site, and are therefore modelled with *adaptive* resolution. Their atomistic or coarse-grained identity is determined by distance from the center of the binding site, such that the atomistic ligand and atomistic protein residues are solvated by a shell of atomistic water at least 1.2 nm thick.

We first describe bonded interactions within the protein, including parametrisation of a coarse-grained protein model and coupling between atomistic and coarse-grained protein. We then outline how non-bonded interactions in the protein and solvent are treated using the AdResS methodology. The model is summarised in Table 1 and illustrated in Figure 1.

2.1 Bonded interactions in the multi-resolution protein model

Proteins may be considered as having both local, high-frequency, small-amplitude fluctuations about conformational substates, and slower, more global transitions between them.^{29,30} This is similar to the “essential dynamics” hypothesis pioneered by Berendsen.³¹ In our minimalistic molecular modelling approach, only those local fluctuations which play a direct role in the biological function of interest are included on an atomistic level, i.e. in this case the ligand binding site. The set of protein degrees of freedom which are modelled atomistically does not change during the simulation, i.e. the protein is fixed-dual-resolution.

The coarse-grained protein exists to ensure the correct structure and conformational fluctuations of the higher resolution binding site. To describe the coarse-grained protein we use an Elastic Network Model¹⁴ in which each residue is mapped to a bead whose location corresponds to the C $_{\alpha}$ atom in the atomistic description. These beads are connected by harmonic springs. The potential energy is then given by

$$E = \sum_i \sum_j k_{ij} (r_{ij} - r_{ij}^0)^2 h(r_c - r_{ij}^0) \quad (1)$$

with spring constants k_{ij} , equilibrium distances r_{ij}^0 , a cutoff distance r_c , and where i, j are nodes and $h(r) = 1$ if $r > 0$ and $h(r) = 0$ otherwise. This family of models has been shown to successfully capture global protein conformational fluctuations including low frequency modes.^{32,33} For the protein used here, lysozyme, this includes the widely studied hinge-bending motion.³⁴ Although the ENM is most often employed for Normal Mode Analysis, it can also be used in molecular dynamics simulations.³⁵

The ENM used here is parametrised such that it reproduces the conformational fluctuations of reference atomistic simulations, as quantified by the root mean square fluctuations (rmsf) of the C $_{\alpha}$ atoms and the NMR S² order parameters of the backbone NH bonds. Order parameters from ¹⁵N spin relaxation experiments are used for cross-validation and B-factors from X-ray crystallography structure determinations are used for comparison. (See the Supporting Material, which also contains some remarks on the relative reliability of different possible sources of reference data for parametrisation). While we primarily use atomistic reference simulations rather than experimental data here for parametrisation, this is only to facilitate the validation process.

The coarse-grained ENM protein is then converted into a dual-resolution model as follows and as illustrated in Figure 1. For an N-residue protein with residues $A = \{A_i, i = 1 \dots N\}$ numbered along the protein backbone, a subset of residues A' are designated as atomistic, and all backbone and sidechain atoms therein are modelled using a standard atomistic forcefield. The residues in A' are not necessarily consecutive in the protein sequence. For a given atomistic residue $A_i \in A'$, if $A_{i-1} \notin A'$ then the backbone C and O atoms in A_{i-1} are also modelled atomistically, and bonded to C_{i-1} using standard atomistic forcefield parameters, with a corresponding reduction in the mass

of the node C_{i-1} , and similarly for the backbone N and H atoms in A_{i+1} if $A_{i+1} \notin A'$. The C_α atoms of the residues in A' remain part of the coarse-grained network. For two C_α atoms with an atomistic bonded interaction, i.e. for consecutive atomistic residues along the protein backbone, the harmonic spring between the corresponding nodes is removed from the ENM. In short, atomistic residues are inserted directly into the ENM, with their C_α atoms replacing ENM nodes. This can be viewed as a step-wise change in resolution, as opposed to the gradual resolution change which takes place across the hybrid region in the solvent. Our procedure allows the coupling of coarse-grained and atomistic protein models without perturbing the structure and conformational fluctuations of either, as will be shown below.

2.2 Non-bonded interactions in the multi-resolution system

To treat the non-bonded interactions, we use the AdResS methodology,²⁴ which was developed to allow the simulation of multi-resolution fluids with free exchange of particles between an atomistic (AT) and a coarse-grained (CG) region. The theoretical basis^{27,36,37} and practical application^{25,38} of this approach have been extensively explored, including its use in the simulation of systems containing fully atomistic biomolecules in solution.^{16,39,40} Here, we extend the methodology, describing how it applies in the case of a dual-resolution non-adaptive macromolecule solvated in adaptive resolution solvent.

A parameter λ designates the resolution of each protein particle or water molecule. We set $\lambda = 1$ for atomistic protein or ligand atoms, and $\lambda = 0$ for coarse-grained protein particles, i.e. each node in the ENM. The λ value for each water molecule depends on its location in space. A spherical region with radius d_{at} is defined centred on a designated atom in the atomistic protein or ligand with coordinates \mathbf{r}_{centre} . Solvent molecules whose centre of mass \mathbf{r}_α satisfies $|\mathbf{r}_{centre} - \mathbf{r}_\alpha| < d_{at}$ have $\lambda = 1$ and are modelled at atomistic resolution. Solvent molecules with $|\mathbf{r}_{centre} - \mathbf{r}_\alpha| > (d_{at} + d_{hy})$ have $\lambda = 0$ and are modelled at a coarse-grained resolution, where each molecule is mapped to a single interaction site at its centre of mass. The quantity d_{hy} is the thickness of a transition or hybrid (HY) region, where the value of λ for solvent molecules varies monotonically from 1 to 0.²⁴

Nonbonded interactions forces are then calculated using the linear force interpolation

$$\mathbf{F}_{\alpha\beta} = \lambda(\mathbf{r}_\alpha)\lambda(\mathbf{r}_\beta)\mathbf{F}_{\alpha\beta}^{AT} + [1 - \lambda(\mathbf{r}_\alpha)\lambda(\mathbf{r}_\beta)]\mathbf{F}_{\alpha\beta}^{CG} \quad (2)$$

where

$$\mathbf{F}_{\alpha\beta}^{AT} = \sum_{i \in \alpha} \sum_{j \in \beta} \mathbf{F}_{ij}^{AT} \quad (3)$$

and α, β are ENM beads or solvent molecule centres of mass, and i, j are atoms in the atomistic protein or in the solvent molecules. In the current work, \mathbf{F}^{AT} means the non-bonded contributions of any standard atomistic protein, ligand and water forcefields (here Amber99SB⁴¹, GLYCAM⁴² and SPCE/E⁴³), and \mathbf{F}^{CG} means any standard coarse-grained model for water-water interactions (here a potential derived via Iterative Boltzmann Inversion^{44,45}) and an excluded-volume interaction between protein and water (see Supporting Material for further details on all potentials). This scheme leads to the following interactions. Non-bonded interactions within the atomistic protein, between two atomistic water molecules, and between atomistic protein and water simplify to \mathbf{F}_{ij}^{AT} . There are no non-bonded interactions between the coarse-grained and atomistic protein, nor within the coarse-grained protein, these interactions being modelled entirely by the ENM. Interactions between two coarse-grained water molecules simplify to $\mathbf{F}_{\alpha\beta}^{CG}$. Interactions within the solvent across the resolution boundaries are treated using the AdResS interpolation (Eq. 2). Interactions between the atomistic protein and hybrid or coarse-grained water also use the interpolation, however in practice d_{at} should be chosen to be large enough so these do not occur. Finally, all water molecules regardless of their resolution have an excluded volume interaction with the coarse-grained protein nodes.

This force-interpolation scheme is inherently non-conservative^{27,37} and a local thermostat must be employed in order to remove the excess heat produced in the hybrid region. An alternative, energy-conserving adaptive resolution scheme exists in which energies rather than forces are interpolated,²⁸ however in this case momentum is no longer strictly conserved. In other words, in the adaptive resolution solvent, one has to choose either energy or momentum conservation. This is a

well known issue,⁴⁶ and does not prevent practical applications. We clarify that this question of choosing between a force-based or energy-based scheme does not arise in the case of the model for bonded interactions in the protein, which is non-adaptive, and is both energy-conserving (that is, a potential energy is well defined) and momentum-conserving.

The AdResS scheme has been demonstrated to yield an atomistic region whose structural and dynamical properties are identical to those of an equivalent region in a much larger fully atomistic system.^{16,24,26,27,39,47} Moreover, this is independent of the quality of the solvent coarse-grained model chosen. In fact, we recently demonstrated that two models as different as atomistic water and a gas of non-interacting particles can be coupled via the AdResS methodology without perturbing the properties of the atomistic water.⁴⁸ One simply needs a reservoir of coarse-grained particles such that particles are correctly supplied to and accepted from the atomistic region.

3 Results and discussion

We performed fully atomistic and multi-resolution molecular dynamics simulations of HEWL, with and without the ligand di-N-acetylchitotriose, an inhibitor. This enzyme hydrolyses glycosidic bonds in polysaccharides. The binding site is in a cleft between two lobes and has both chemical and steric specificity, both of which our model will reproduce. In the multi-resolution enzyme, the residues modelled on an atomistic level are those that form stable H-bonds to the ligand (Asn-59, Trp-62, Trp-63 and Ala-107) and the four nearest neighbours in terms of the distance of their centre of mass from the ligand (Ile-58, Ile-98, Asp-101, Trp-108, see Supporting Material). This is illustrated in Figure 2.

For the study of processes such as inhibitor binding or enzyme catalysis using a multi-resolution model, the ligand must experience an environment as close as possible to the true environment. Here, we examine the properties of the ligand and the binding site in the multi-resolution model, comparing them to those determined experimentally or via fully atomistic simulations. In this context, our current understanding of the biological process of interest must be borne in mind. For example, ligand binding may be understood via the opposing induced-fit or population-shift models.⁴⁹ Similarly, for enzyme catalysis, opposing hypotheses posit either stabilisation of the transition state by a “pre-organised” environment, or coupling of certain vibrational modes in the enzyme to the enzymatic reaction coordinate.^{10,11} Such open questions can be seen as a challenge to be overcome in the construction of a multi-resolution model, or rather as an opportunity for multi-resolution models to contribute to the debate. As a first step, the multi-resolution model must reproduce the relevant enzyme properties in all cases. We first discuss the global properties of the protein, before turning to the properties of the binding site.

3.1 Global protein structure and fluctuations

The protein’s global properties determine the local properties of the binding site. In the ENM, the global structure of the protein is assured by construction. As an aside, although an ENM has only one equilibrium structure, proteins with two or more distinct and well-defined conformational states on a global level could be represented using a double or multi-well ENM.⁵⁰ The global conformational fluctuations of the coarse-grained model, controlled by k_{ij} , were parametrised in a single-resolution, fully coarse-grained ENM in the first step of model-building, and these parameters were then used in the multi-resolution model. Figure 3 shows the rmsf of the ENM beads and C_α atoms in the multi-resolution model versus the rmsf of the ENM beads in the single-resolution ENM. The global fluctuations of the protein, as parametrised in the purely coarse-grained model, are clearly not perturbed by the insertion of atomistic detail. We note that the rmsf values of the C_α atoms in the multi-resolution model, marked by black arrows, cannot be compared directly to the rmsf values of the corresponding ENM nodes in the single-resolution model, as these represent the movement of an entire residue.

3.2 Properties of the binding site in the ligand-free system

We now turn to key properties of the binding site, in each case comparing multi-resolution and fully atomistic simulations.

Starting with the ligand-free system, we study the structure and conformational fluctuations of the binding site via the distributions of distances between the four C_α atoms in the core H-bonding residues (Asn-59, Trp-62, Trp-63 and Ala-107). These are shown in Figure 4.

Residues Asn-59, Trp-62 and Trp-63 are in the enzyme’s β -lobe, while Ala-107 is in the α -lobe. The first three distances shown (Ala107:CA-Asn59:CA, Ala107:CA-Trp62:CA and Ala107:CA-Trp63:CA, Figure 4(a-c)) cross the binding cleft. The agreement between atomistic and multi-resolution structures, as measured by the distributions’ maxima, is very good; however, the distributions in the multi-resolution system are considerably narrower. Here, we see a result of the harmonic approximation inherent in the ENM. The protein fluctuations as described by the fully atomistic forcefield have a much greater variance, and while the distributions examined here remain reasonably symmetric and monomodal, this will not necessarily be the case for fluctuations across the active site of all proteins. The remaining two distances shown (Figure 4(d,e)) are within the β -lobe. Again, the agreement between atomistic and multi-resolution simulations is good, within the limits of the harmonic ENM model. The ENM used here has only two spring constants, k_b for $i...i + 1$ interactions along the backbone and k_{nb} for all other interactions. The agreement between multi-resolution and atomistic fluctuations could of course be improved by the use of a more complex ENM model, or at least by tuning specific spring constants, for example for $i...i + 2$ and $i...i + 3$ interactions along the backbone. However such detailed information on the local conformational fluctuations is not available experimentally. Instead of overfitting our model to perhaps unreliable atomistic simulations, we prefer to retain simplicity, and therefore applicability to a wide range of systems. This simple but widely used ENM is well known to successfully model global protein fluctuations.¹⁵ In the context of our multi-resolution model, it supplies the global structure and fluctuations that determine the local properties of the binding site, allowing modelling of the phenomenon of molecular recognition, as shown below.

A unique feature of our methodology is its incorporation of the AdResS treatment of the solvent, allowing us to solvate the atomistic binding site in a small sphere of atomistic water, which is in turn immersed in a less expensive coarse-grained solvent, but which behaves as though it were in a fully atomistic solvent. The sphere is centred on an atom of the protein or ligand (see Supporting Material) and its radius d_{at} is 2.4 nm, such that there is always at least 1.2 nm between any atomistic protein or ligand atom and the atomistic/hybrid boundary. A large portion of this sphere is occupied by the excluded volume of the protein, and it contains on average 1450 fully atomistic water molecules. This number rises to 3250 if we include all water molecules with $\lambda > 0.5$, i.e. with at least 50 % atomistic character. To confirm that this is sufficient to reproduce the behaviour of a fully atomistic simulation, we now examine the properties of water in the system and the hydration of the ligand-free binding site.

We first calculate the water density across the simulation box. This is done by dividing the box into a three-dimensional grid of subcells, assigning subcells to the AT, HY and CG region at each timestep, excluding those subcells which overlap two regions or contain protein excluded volume, and then averaging the recorded values in time for each region. The bulk water density is 0.995 ± 0.002 , 1.001 ± 0.001 and 0.996 ± 0.001 g cm⁻³ in the AT, HY and CG regions respectively. This is in excellent agreement with the reference value of 1.000 ± 0.001 g cm⁻³ from the fully atomistic simulations of the same system.

We then examine the hydration of the binding site. We identify the four atoms which form stable H-bonds with the inhibitor when it is present: Asn-59:H, Trp-62:HE1, Trp-63:HE1 and Ala-107:O. In the inhibitor-free system, we then calculate the average number of hydrogen bonds to water formed at those key sites, i.e. these are H-bonds to the water molecules which are displaced by ligand binding. Figure 5 shows the comparison between fully atomistic and multi-resolution simulations. The agreement is excellent for the three donor and one acceptor sites. We note that the correct modelling of the binding site hydration is made possible by the AdResS treatment of the solvent, which ensures that the density and other properties of water are correctly modelled even with only a finite sphere of atomistic water around the binding site. While we do not perform an exhaustive study of water properties here, it has been shown elsewhere that the structure and dynamics of water molecules hydrating protein surfaces can be well reproduced with the AdResS setup.¹⁶

3.3 Properties of the binding site in the ligand-bound system

We now turn to the interactions between the inhibitor and the protein. Figure 6 shows the probability density distributions for the key distances which quantify contacts between the binding site and the inhibitor. This includes two distances measuring the hydrophobic contact between the Trp-62 aromatic sidechain and the ligand (Figure 6(a,b)) and the heavy atom-hydrogen distances in five H-bonds (to Asn-59:H, Trp-62:HE1, Trp-63:HE1 twice, and Ala-107:O, Figure 6(c-g)). The inhibitor clearly remains stably bound to the dual-resolution protein over the entire 14-ns simulation, forming inhibitor-protein contacts exactly as in the fully atomistic model. For the one case where the comparison is good but not perfect (Trp-62:HE1–NAG:O6A, Figure 6(d)), the H-bond is weak, and is only formed a small portion of the time. While the H-bond is broken, this distance samples a large conformational space, making it more difficult to reproduce.

According to the electrostatic stabilisation hypothesis,¹¹ the electric field of the protein plays a key role in enzyme catalysis by stabilising the transition state of the reaction. Studies from the field of non-aqueous enzymology suggest that the polarity of the aqueous solvent also plays a role. In Figure 7, we show the magnitude of the electric field from the protein and from water on key atoms in the ligand, in the fully atomistic and multi-resolution simulations. In the same figure we also show the electrostatic potential in the binding site, calculated on the surface of a cylinder enclosing the ligand.⁵¹ The potential at each point in the plane is calculated as a summation over atomistic charges, in order to allow a comparison between the particle-based fully atomistic and multi-resolution models. The electrostatic potential shown is that due to atomistic protein and solvent charges, excluding charges of the ligand, i.e. that felt by the ligand due to its environment. Again, the agreement is good, completing the demonstration that the environment felt by the ligand in the multi-resolution model matches that in a fully atomistic model. Were the multi-resolution model used to study enzymatic catalysis, the transition-state stabilisation effect would be satisfactorily included.

The analysis outlined above demonstrates that our model can accurately capture the interaction of a ligand with the protein’s binding site. To do so, a sufficient but still very small number of residues must be included in atomistic detail. The eight atomistic residues used here contain only 8.5% of the total atomistic degrees of freedom in the protein. The precise details of this will vary from enzyme to enzyme and will form the subject of a future work.

4 Conclusion

We have developed a multi-resolution methodology in which arbitrary degrees of freedom in a protein/water system can be included in atomistic detail. The multi-resolution model captures the global properties of the protein via a highly coarse-grained model, while still allowing for chemical detail in the ligand binding site. This highly minimalistic model is enough to model stable ligand binding with the inclusion on an atomistic level of only 8.5% of the enzyme’s total atomistic degrees of freedom. For larger proteins this proportion is expected to be even lower.

Our methodology is transferable to any atomistic protein and solvent force field. As discussed earlier, the choice of the coarse-grained solvent force field has no effect on the properties of the atomistic region. Regarding the coarse-grained protein model, the ENM used here could be replaced by a very different ansatz, such as a Gō-like model or multi-bead protein model,⁵² in order to capture different protein features, for example large-scale conformational change or side-chain flexibility. Coarse-graining always represents a compromise, and the choice of coarse-grained protein model will be dictated by the biophysical or biochemical processes being studied in any given application.

Our approach to the coarse-grained protein allows a parametrisation directly on experiment, instead of using as a reference a possibly problematic atomistic forcefield. Although atomistic protein force fields form an essential part of the biomolecular simulation toolkit, they also have well-known weaknesses, and can deviate from the experimentally observed behaviour with increasing trajectory length, due to the accumulation of errors in the parametrization of the force field.^{53–55} In our approach, the use of the atomistic force field is limited to the binding site region. Moreover, this allows the simulation of systems for which the structure is known in high resolution only in some parts of the system, as well as systems that are too large for a fully atomistic modelling.

This opens the way to a range of applications. In structure-based drug design, for example,

docking or molecular recognition studies must use simulation methods which are as efficient as possible, due to the large number of lead compounds which must be screened.⁵⁶ The receptor is usually treated as a rigid body for reasons of computational cost, however it has been shown that receptor flexibility and thermal motions play an important role in molecular recognition.^{56,57} The model presented here could be used to combine both flexibility and computational efficiency.

Beyond computational efficiency, a model which allows the inclusion and exclusion of arbitrary degrees of freedom while still correctly modelling structure, dynamics and thermodynamics can be used to pinpoint and quantify the contribution of any given set of degrees of freedom to biophysical or biochemical processes in the system, for example free energies of binding or enzymatic reaction rate constants. Enzymatic catalysis could be studied using the current purely classical methodology via the Empirical Valence Bond method,⁵⁸ without needing the inclusion of a quantum/classical coupling in addition to the classical atomistic/coarse-grained coupling.

Future refinement to the multi-resolution model presented here will include a coarse-grained model capable of capturing anharmonic fluctuations, and inclusion of counterions in the AdResS treatment of the solvent.

5 Author contributions

A.C.F., R.P. and K.K. designed research; A.C.F. performed research; A.C.F., R.P. and K.K. wrote the manuscript.

6 Supporting Material

Additional supplemental information including three figures, parametrisation of the coarse-grained models, and simulation details are available online at <http://onlinelibrary.wiley.com/doi/10.1002/prot.25173/full>.

7 Acknowledgements

K.K. and A.C.F. acknowledge research funding through the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. 340906-MOLPROCOMP. We are grateful to Torsten Stuehn for assistance with the ESPResSo++ package and to Debashish Mukherji and Tristan Bereau for a critical reading of the manuscript.

References

- [1] O. B. Okan, A. R. Atilgan and C. Atilgan, *Biophys. J.*, 2009, **97**, 2080 – 2088.
- [2] M. Karplus and J. A. McCammon, *Annu. Rev. Biochem.*, 1983, **53**, 263–300.
- [3] K. A. Dill and J. L. MacCallum, *Science*, 2012, **338**, 1042–1046.
- [4] K. Meier, A. Choutko, J. Dolenc, A. P. Eichenberger, S. Riniker and W. F. van Gunsteren, *Angew. Chem. Int. Ed.*, 2013, **52**, 2820–2834.
- [5] V. Tozzini, *Acc. Chem. Res.*, 2010, **43**, 220–230.
- [6] T. A. Wassenaar, H. I. Ingólfsson, M. Prie, S. J. Marrink and L. V. Schäfer, *J. Phys. Chem. B*, 2013, **117**, 3516–3530.
- [7] S. Riniker, A. P. Eichenberger and W. F. van Gunsteren, *Eur. Biophys. J.*, 2012, **41**, 647–661.
- [8] Q. Shi, S. Izvekov and G. A. Voth, *J. Phys. Chem. B*, 2006, **110**, 15045–15048.
- [9] J. K. Sigurdsson, F. L. Brown and P. J. Atzberger, *J. Comput. Phys.*, 2013, **252**, 65 – 85.
- [10] S. C. L. Kamerlin and A. Warshel, *Proteins: Struct., Funct., Bioinf.*, 2010, **78**, 1339–1375.
- [11] M. Garcia-Viloca, J. Gao, M. Karplus and D. G. Truhlar, *Science*, 2004, **303**, 186–195.

- [12] R. Affleck, Z. F. Xu, V. Suzawa, K. Focht, D. S. Clark and J. S. Dordick, *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 1100–1104.
- [13] A. M. Klibanov, *Trends Biotechnol.*, 1997, **15**, 97–101.
- [14] M. M. Tirion, *Phys. Rev. Lett.*, 1996, **77**, 1905–1908.
- [15] T. D. Romo and A. Grossfield, *Proteins: Struct., Funct., Bioinf.*, 2011, **79**, 23–34.
- [16] A. C. Fogarty, R. Potestio and K. Kremer, *J. Chem. Phys.*, 2015, **142**, 195101.
- [17] O. M. Szklarczyk, N. S. Bieler, P. H. Hünenberger and W. F. van Gunsteren, *J. Chem. Theory Comput.*, 2015, **11**, 5447–5463.
- [18] N. di Pasquale, D. Marchisio and P. Carbone, *J. Chem. Phys.*, 2012, **137**, 164111.
- [19] C. F. Abrams, L. Delle Site and K. Kremer, *Phys. Rev. E*, 2003, **67**, 021807.
- [20] M. Neri, C. Anselmi, M. Cascella, A. Maritan and P. Carloni, *Phys. Rev. Lett.*, 2005, **95**, 218102.
- [21] M. Neri, M. Baaden, V. Carnevale, C. Anselmi, A. Maritan and P. Carloni, *Biophys. J.*, 2008, **94**, 71–78.
- [22] M. R. Machado, P. D. Dans and S. Pantano, *Phys. Chem. Chem. Phys.*, 2011, **13**, 18134–18144.
- [23] M. R. Machado and S. Pantano, *J. Chem. Theory Comput.*, 2015, **11**, 5012–5023.
- [24] M. Praprotnik, L. Delle Site and K. Kremer, *J. Chem. Phys.*, 2005, **123**, 224106.
- [25] M. Praprotnik, S. Matysiak, L. Delle Site, K. Kremer and C. Clementi, *J. Phys. Condens. Matt.*, 2007, **19**, 292201.
- [26] S. Matysiak, C. Clementi, M. Praprotnik, K. Kremer and L. Delle Site, *J. Chem. Phys.*, 2008, **128**, 024503.
- [27] H. Wang, C. Hartmann, C. Schütte and L. Delle Site, *Phys. Rev. X*, 2013, **3**, 011018.
- [28] R. Potestio, S. Fritsch, P. Español, R. Delgado-Buscalioni, K. Kremer, R. Everaers and D. Donadio, *Phys. Rev. Lett.*, 2013, **110**, 108301.
- [29] C. Baysal and A. R. Atilgan, *Biophys. J.*, 2002, **83**, 699–705.
- [30] H. Frauenfelder and B. McMahon, *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 4795–4797.
- [31] A. Amadei, A. B. M. Linssen and H. J. C. Berendsen, *Proteins: Struct., Funct., Bioinf.*, 1993, **17**, 412–425.
- [32] N. Leioatts, T. D. Romo and A. Grossfield, *J. Chem. Theory Comput.*, 2012, **8**, 2424–2434.
- [33] L. Orellana, M. Rueda, C. Ferrer-Costa, J. R. Lopez-Blanco, P. Chacón and M. Orozco, *J. Chem. Theory Comput.*, 2010, **6**, 2910–2923.
- [34] K. N. Woods, *J. Biol. Phys.*, 2014, **40**, 121–137.
- [35] W. Zheng and P. Glenn, *J. Chem. Phys.*, 2015, **142**, 035101.
- [36] S. Fritsch, S. Poblete, C. Junghans, G. Ciccotti, L. Delle Site and K. Kremer, *Phys. Rev. Lett.*, 2012, **108**, 170602.
- [37] A. Agarwal, H. Wang, C. Schütte and L. Delle Site, *J. Chem. Phys.*, 2014, **141**, 034102.
- [38] M. Praprotnik, L. Delle Site and K. Kremer, *J. Chem. Phys.*, 2007, **126**, 134902.

- [39] J. Zavadlav, M. N. Melo, S. J. Marrink and M. Praprotnik, *J. Chem. Phys.*, 2014, **140**, 054114.
- [40] D. Mukherji, N. F. A. van der Vegt and K. Kremer, *J. Chem. Theory Comput.*, 2012, **8**, 3536–3541.
- [41] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins: Struct., Funct., Bioinf.*, 2006, **65**, 712–725.
- [42] K. N. Kirschner, A. B. Yongye, S. M. Tschampel, J. González-Outeiriño, C. R. Daniels, B. L. Foley and R. J. Woods, *J. Comput. Chem.*, 2008, **29**, 622–655.
- [43] H. Berendsen, J. Grigera and T. Straatsma, *J. Phys. Chem.*, 1987, **91**, 6269–6271.
- [44] A. Soper, *Chem. Phys.*, 1996, **202**, 295–306.
- [45] D. Reith, M. Pütz and F. Müller-Plathe, *J Comput. Chem.*, 2003, **24**, 1624–1636.
- [46] L. Delle Site, *Phys. Rev. E*, 2007, **76**, 047701.
- [47] F. Stanzione and A. Jayaraman, *J. Phys. Chem. B*, 2016, **120**, 4160–4173.
- [48] K. Kreis, A. C. Fogarty, K. Kremer and R. Potestio, *Eur. Phys. J. Spec. Top.*, 2015, **224**, 2289–2304.
- [49] K.-i. Okazaki and S. Takada, *Proc. Natl. Acad. Sci. USA*, 2008, **105**, 11182–11187.
- [50] J.-W. Chu and G. A. Voth, *Biophys. J.*, 2007, **93**, 3860–3871.
- [51] For the electrostatic potential calculations, a coordinate system was defined in the frame of reference of the ligand. Its origin was at the ligand centre of mass, and its axes were the first three principal components of the covariance matrix of a set of four ligand heavy atoms in the sugar rings.
- [52] V. Tozzini, *Curr. Opin. Struct. Biol.*, 2005, **15**, 144–150.
- [53] D. Petrov and B. Zagrovic, *PLoS Comput. Biol.*, 2014, **10**, 1–11.
- [54] S. Piana, J. L. Klepeis and D. E. Shaw, *Curr. Opin. Struct. Biol.*, 2014, **24**, 98–105.
- [55] P. L. Freddolino, S. Park, B. Roux and K. Schulten, *Biophys. J.*, 2009, **96**, 3772–3780.
- [56] N. Kamiya, Y. Yonezawa, H. Nakamura and J. Higo, *Proteins: Struct., Funct., Bioinf.*, 2008, **70**, 41–53.
- [57] K. W. Lexa and H. A. Carlson, *J. Am. Chem. Soc.*, 2011, **133**, 200–202.
- [58] S. C. L. Kamerlin and A. Warshel, *Faraday Discuss.*, 2010, **145**, 71–106.

	AT protein	ENM protein	AT water	HY water	CG water
AT protein	F_{AT}				
ENM protein	F_{ENM}	F_{ENM}			
AT water	F_{AT}	F_{WCA}	F_{AT}		
HY water	$\lambda F_{AT} + (1 - \lambda)F_{WCA}$ [1]	F_{WCA}	$\lambda F_{AT} + (1 - \lambda)F_{IBI}$	$\lambda F_{AT} + (1 - \lambda)F_{IBI}$	
CG water	F_{WCA} [1]	F_{WCA}	- [2]	F_{IBI}	F_{IBI}
[1] not used if d_{at} is large enough, [2] because $d_{hy} \geq$ non-bonded cutoff					

Table 1: Summary of interactions in the multi-resolution model. F_{ENM} = Elastic Network Model; F_{AT} = atomistic AMBER+GLYCAM forcefield and SPC/E water; $F_{CG} = F_{WCA}, F_{IBI}$, where F_{WCA} = WCA (excluded volume) interaction and F_{IBI} = potential from IBI coarse-graining of atomistic (SPC/E) water; AT protein: $\lambda = 1$, ENM protein: $\lambda = 0$, water: $0 < \lambda < 1$

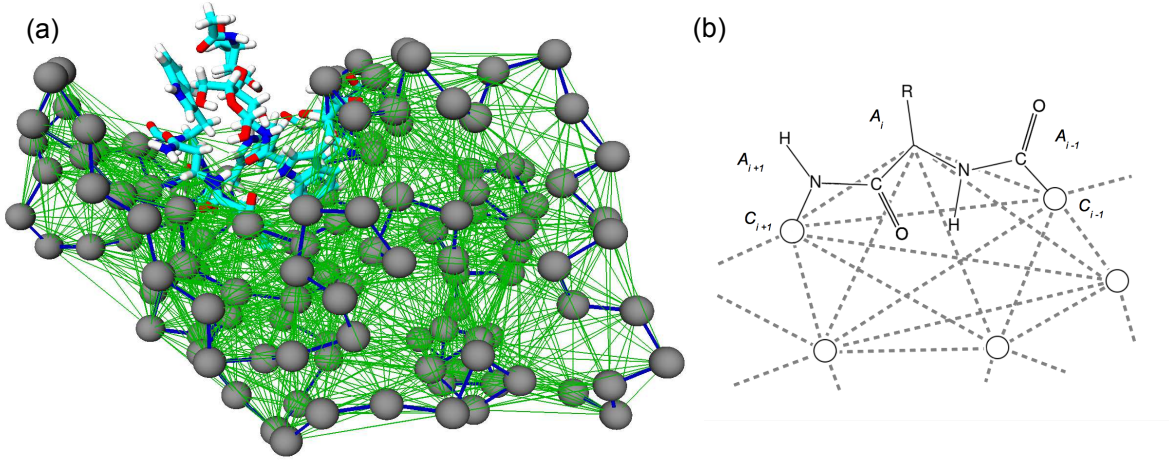


Figure 1: (a) Visualisation of the multi-resolution protein model. The residues included in atomistic detail are shown in red, blue, cyan and white (O, N, C and H atoms). The grey spheres are ENM nodes, the stiff backbone springs are shown as dark blue lines and all other (weaker) springs are shown in green. (b) Bonded coupling between an atomistic protein residue and the coarse-grained protein model. Black lines and letters are bonds and atoms described using the atomistic forcefield, with 'R' representing any sidechain, circles are ENM nodes and dashed grey lines are ENM springs.

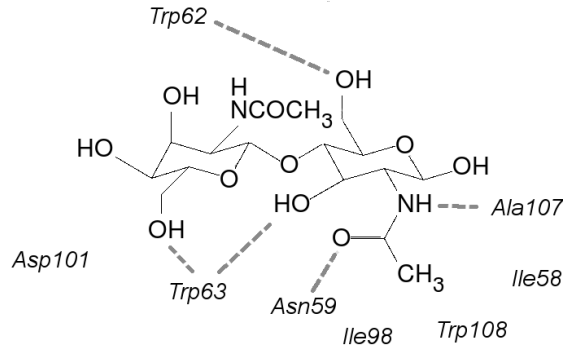


Figure 2: Atomistic residues and inhibitor in the binding site of the protein. Dashed grey lines are hydrogen bonds.

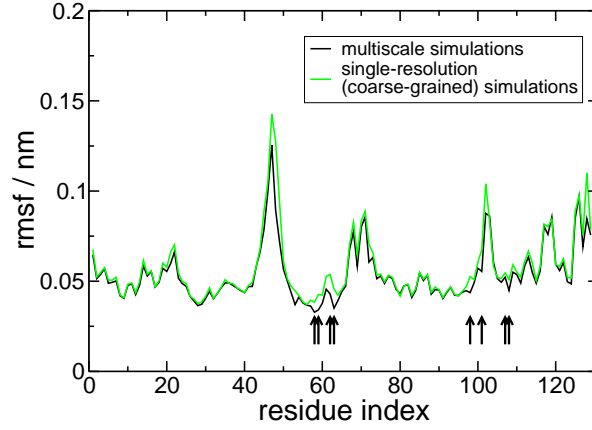


Figure 3: Rmsf of beads in the single-resolution ENM in single-resolution solvent vs rmsf of ENM beads and protein C_α atoms in the multi-resolution system. The black arrows mark the positions of the atomistic residues in the multi-resolution case.

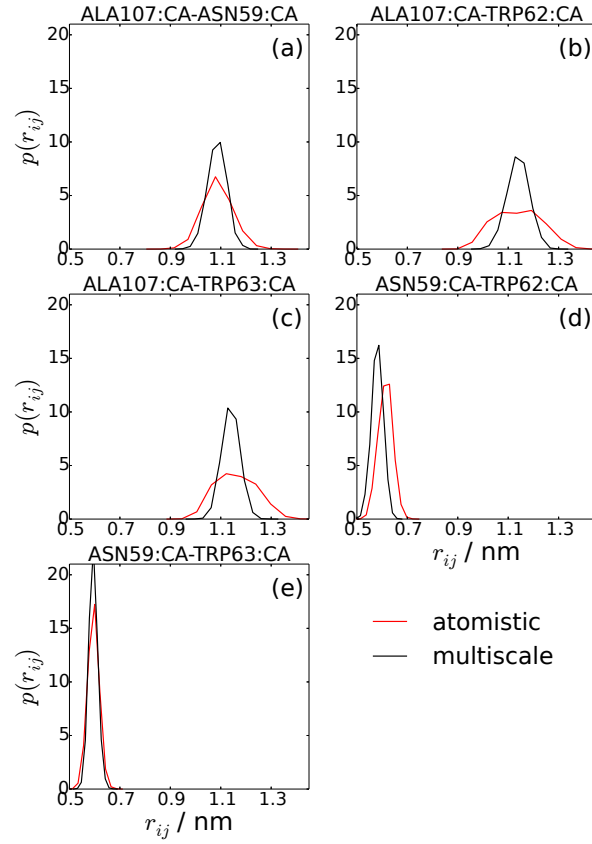


Figure 4: Probability density distributions for key protein-protein distances across the binding site in the ligand-free system, multi-resolution versus fully atomistic simulations.

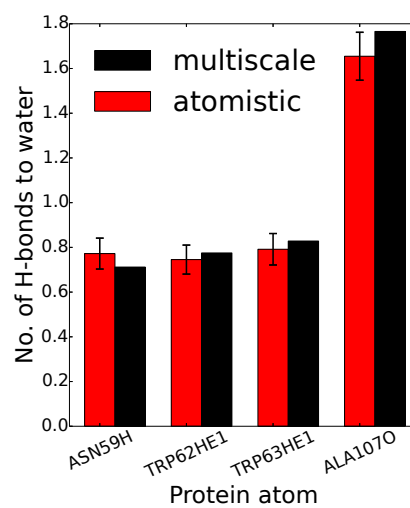


Figure 5: Number of H-bonds between water and the key H-bond acceptors and donors in the binding site, in the ligand-free system, multi-resolution versus fully atomistic simulations.

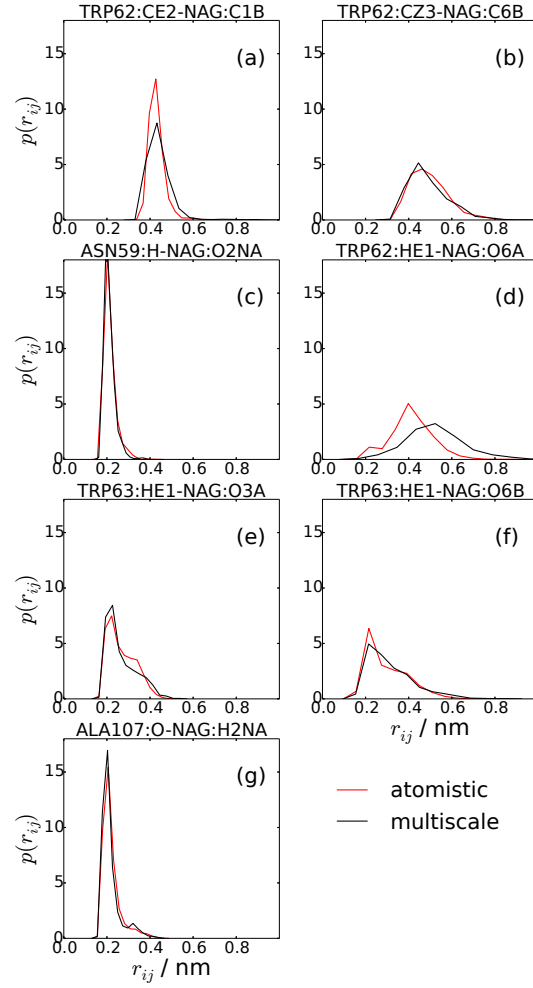


Figure 6: Probability density distributions for the key distances in the hydrophobic contacts (a,b) and the H-bonds (c-g) between the protein and the ligand, multi-resolution versus fully atomistic simulations.

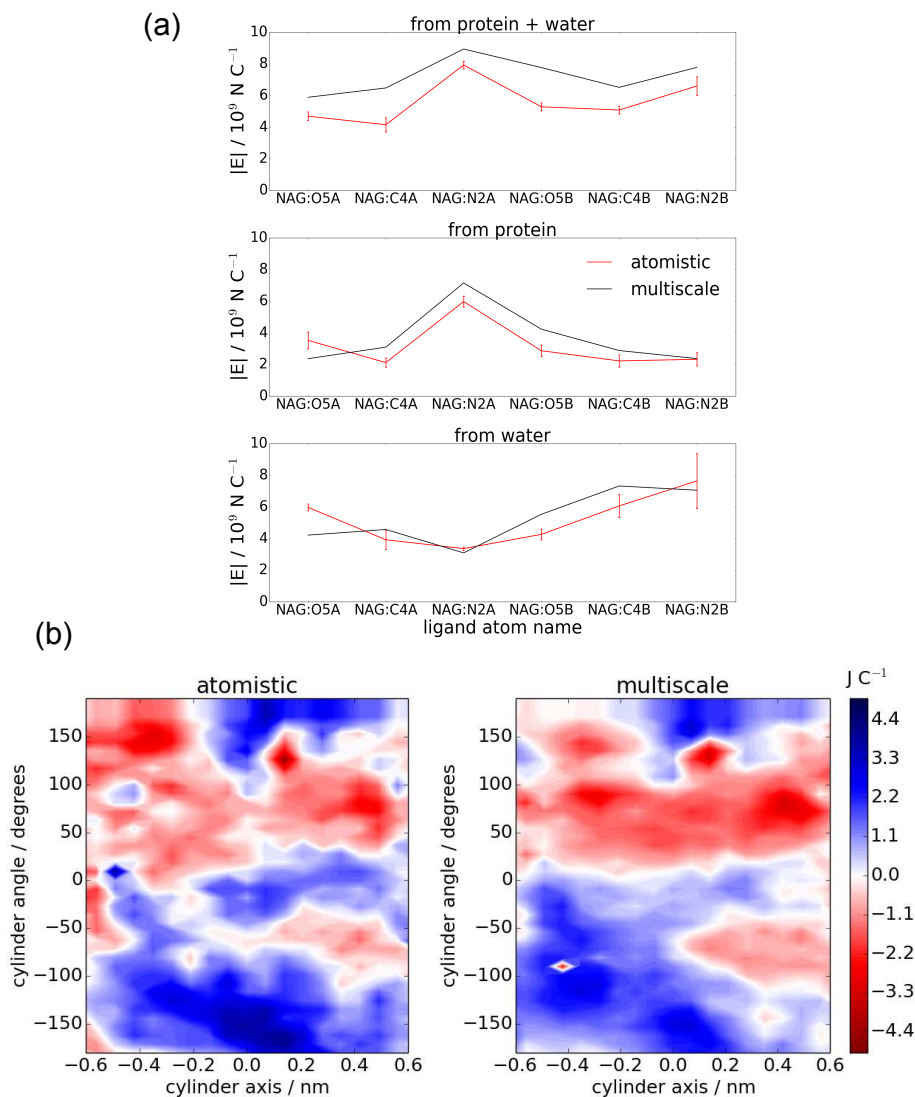


Figure 7: (a) Electric field on selected atoms in the ligand, from the protein and from water, multi-resolution versus fully atomistic simulations. (b) Electrostatic potential felt by the ligand in the binding site due to protein and solvent charges, on the surface of a cylinder of radius 0.3 nm enclosing the ligand. The cylinder is centered on the ligand center of mass and its axis runs along the ligand's longest dimension, such that the cylinder's size, shape and orientation match that of the ligand.